Untersuchungen zur koronaren Herzkrankheit (t-tests)

Untersuchungen zur koronaren Herzkrankheit

In diesem Abschnitt sollen Daten von Probanden bzw. Patienten auf das Risiko für koronare Herzkrankheit untersucht werden. Dies ist eine Erkrankung der Herzkranzgefäße (Koronararterien), die sich durch Ablagerungen in den Gefäßwänden verengen. Der Original-Herz-Datensatz ist unter

- https://archive.ics.uci.edu/ml/datasets/Heart+Disease

beschrieben. Wir nutzen eine konsolidierte CSV-Datei, die bereits Header enthält. Download unter:

- https://oc.informatik.hs-mannheim.de/s/wyzFq34K9HiNjXR/download

Die Datei enthält 13 Merkmale, die einen Einfluss auf eine koronare Herzkrankheit haben können. Das 14. Merkmal `goal` (im Original auch `num`) ist die Diagnose (Klassifizierung). Der Wert ist 0, falls keine krankhafte Verengung der Gefäße vorliegt, oder 1, 2, 3 oder 4, falls – je nach Stärke – eine krankhafte Verengung der Gefäße vorliegt. Wir unterscheiden im Folgenden nur die Zustände "gesund" (0) und "krank" (1, 2, 3, 4). Unter

- https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/heart-disease.names

finden Sie eine Beschreibung aller Attribute. Hier ist eine Zusammenfassung. Wir benötigen insbesondere die Merkmale `sex`, `trestbps`, `chol` und `goal`.

Feld	Bedeutung
age	age in years
sex	sex (1 = male; 0 = female)
cp	chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic)
trestbps	resting systolic blood pressure (in mmHg on admission to the hospital)
chol	serum cholestoral in mg/dl
fbs	fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
restecg	resting electrocardiographic results (0 = normal; 1 = having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV); 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)
thalach	maximum heart rate achieved
exang	exercise induced angina (1 = yes; 0 = no)
oldpeak	ST depression induced by exercise relative to rest.[1]
slope	slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 = downsloping)
ca	number of major vessels (0-3) colored by flourosopy
thal	3 = normal; 6 = fixed defect; 7 = reversable defect

[1] ST depression refers to a finding on an electrocardiogram, wherein the trace in the ST segment is abnormally low below the baseline.

Feld	Bedeutung
goal	diagnosis of heart disease (0: $< 50\%$ diameter narrowing ; 1, 2, 3, 4: $> 50\%$ diameter narrowing)

Einlesen der Herz-Daten

Lesen Sie die Datei aus der URL als Data Frame zur weiteren Bearbeitung ein. Überlegen Sie, ob sie Faktoren sinnvoll einsetzen können. Geben Sie die ersten drei Zeilen und fünf Spalten aus[2]:

```
#Einlesen der Datei
herz = read.csv(url("https://oc.informatik.hs-mannheim.de/s/wyzFq34K9HiNjXR/download"))

herz$sex = factor(herz$sex, levels = c(0, 1), labels = c("f", "m"))
herz$goal = factor(herz$goal, levels = c(0, 1, 2, 3, 4), labels = c("gesund", "krank", "krank", "krank"

# Ausgabe der ersten 3 Zeilen und 5 Spalten

head(herz, n = c(3, 5))
```

```
##   age sex cp trestbps chol
## 1  63   m  1      145  233
## 2  67   m  4      160  286
## 3  67   m  4      120  229
```

Cholesterin im Vergleich Männer/Frauen

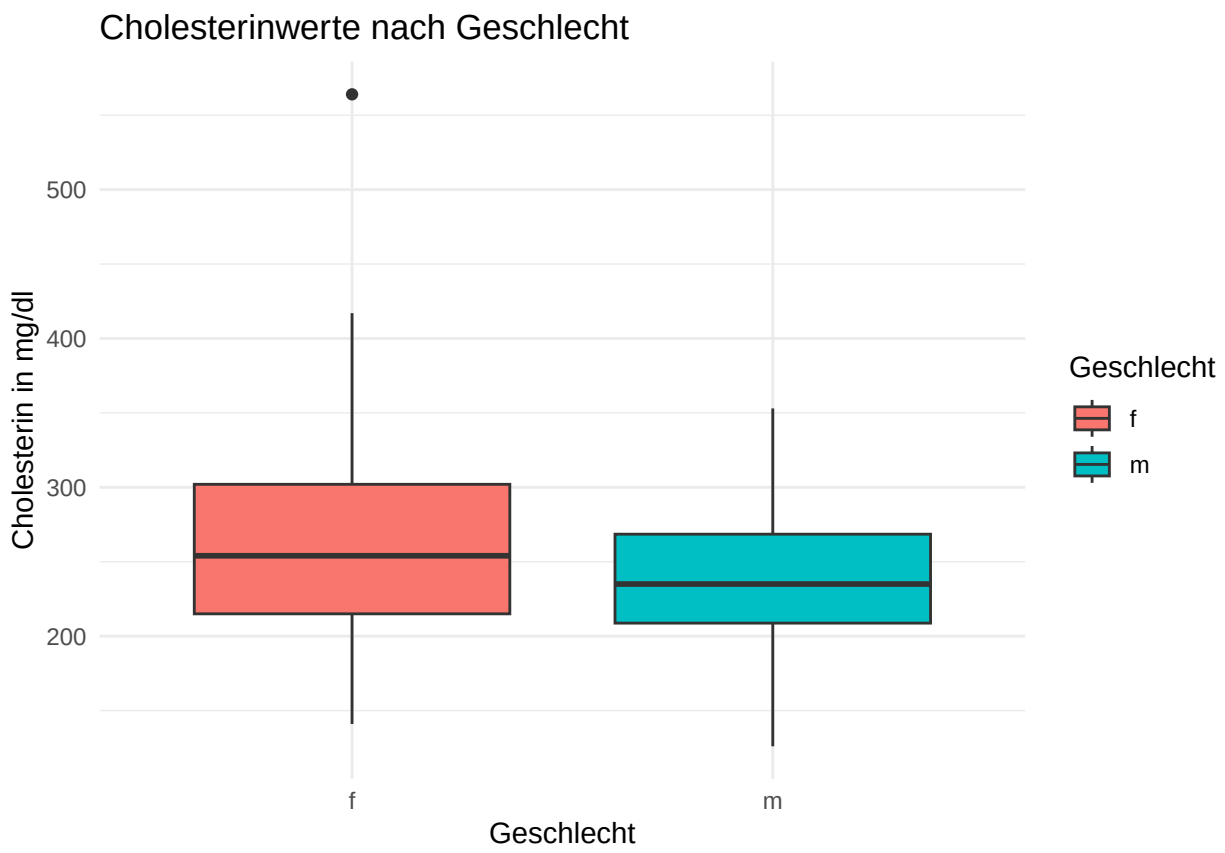
Nun sollen die Cholerstin-Werte untersucht werden – zunächst im Vergleich Männer zu Frauen.

Überblick über Cholesterin-Daten

Verschaffen Sie sich einen Überblick, indem Sie ein Boxplot für das Cholesterin gruppiert nach dem Geschlecht plotten.

```
ggplot(herz, aes(x = sex, y = chol, fill = sex)) +
  geom_boxplot() +
  labs(title = "Cholesterinwerte nach Geschlecht",
       x = "Geschlecht",
       y = "Cholesterin in mg/dl") +
  scale_fill_discrete(name = "Geschlecht") +
  theme_minimal()
```

[2] Möglicherweise kommt es zu einem Fehler beim Einlesen des ersten Attributs (`age`). Manuelles Umbenennen hilft.



Konfidenz-Intervall

Berechnen Sie das Konfidenz-Intervall (Niveau 95%) für den Cholesterin-Level jeweils für Männer und Frauen.

Tabelle Geben Sie das Ergebnis als `kable`-Tabelle aus:

```
# Filtern nach Geschlecht & berechnen des Konfidenzintervalls

konf_chol_men = t.test(herz[herz$sex == "m", ]$chol, conf.level = 0.95)$conf.int
konf_chol_women = t.test(herz[herz$sex == "f", ]$chol, conf.level = 0.95)$conf.int

# Erstellen der Tabelle
result_table_men_vs_women = data.frame(Geschlecht = c("Männer", "Frauen"),
                          Untere_Grenze = c(konf_chol_men[1], konf_chol_women[1]),
                          Obere_Grenze = c(konf_chol_men[2], konf_chol_women[2])
                          )
kable(result_table_men_vs_women)
```

Geschlecht	Untere_Grenze	Obere_Grenze
Männer	233.7432	245.4607
Frauen	248.6722	274.8330

Überlappung? Überlappen sich die Bereiche?

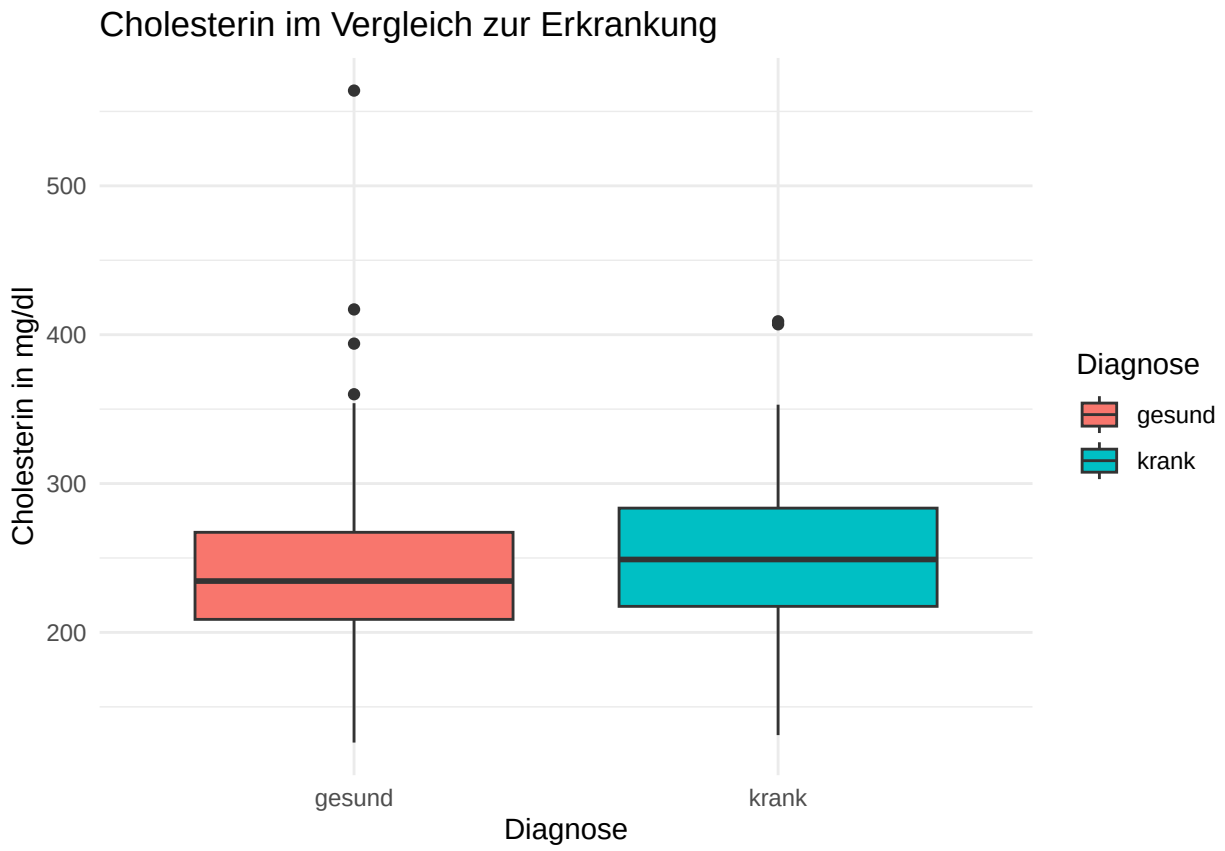
Cholesterin im Vergleich zur Erkrankung

Nun sollen die Cholesterin-Werte in Abhängigkeit der Diagnose untersucht werden.

Überblick über Cholesterin-Daten

Verschaffen Sie sich einen Überblick, indem Sie ein Boxplot für das Cholesterin gruppiert nach der Diagnose plotten.

```
ggplot(herz, aes(x = goal, y = chol, fill = goal)) +
  geom_boxplot() +
  labs(title = "Cholesterin im Vergleich zur Erkrankung",
       x = "Diagnose",
       y = "Cholesterin in mg/dl") +
  scale_fill_discrete(name = "Diagnose") +
  theme_minimal()
```



Konfidenz-Intervall

Berechnen Sie die Konfidenzintervalle für beide Gruppen und geben Sie das Ergebnis als `kable`-Tabelle aus:

```
konf_gesund = t.test(herz[herz$goal == "gesund", ]$chol, conf.level = 0.95)$conf.int
konf_krank = t.test(herz[herz$goal == "krank", ]$chol, conf.level = 0.95)$conf.int
```

```r
result_table_gesund_vs_krank = data.frame(Diagnose = c("Gesund","Krank"),
                                          Untere_Grenze = c(konf_gesund[1], konf_krank[1]),
                                          Obere_Grenze = c(konf_gesund[2], konf_krank[2])
                                          )
kable(result_table_gesund_vs_krank)
```

Diagnose	Untere_Grenze	Obere_Grenze
Gesund	234.3977	250.8828
Krank	243.1752	259.7744

```
### Test

Es sieht so aus, als ob der Cholesterin-Wert bei den erkrankten Patienten höher ist als bei den nicht e

#### Wie lauten die Hypothesen?

Formulieren Sie die Hypothesen ($H_0$ und $H_1$).


#### Testanwendung

Wenden Sie den Test mit R an. Was ist das Ergebnis?


'''r
# Nullhypothese (H0) = Cholesterinwert bei Kranken ist gleich oder kleiner wie bei Gesunden
# Alternativhypothese (H1) Cholesterinwert bei Kranken ist höher als bei Gesunden


test_diagnose = t.test(herz$chol[herz$goal == "krank"],
                       herz$chol[herz$goal == "gesund"],
                       alternative = "greater"
                       ) # betrachten das der Cholesterinwert bei Kranken höher ist als bei Gesunden

print(test_diagnose)

##
##  Welch Two Sample t-test
##
## data:  herz$chol[herz$goal == "krank"] and herz$chol[herz$goal == "gesund"]
## t = 1.4924, df = 298.64, p-value = 0.06832
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.9327489        Inf
## sample estimates:
## mean of x mean of y
##  251.4748  242.6402
```

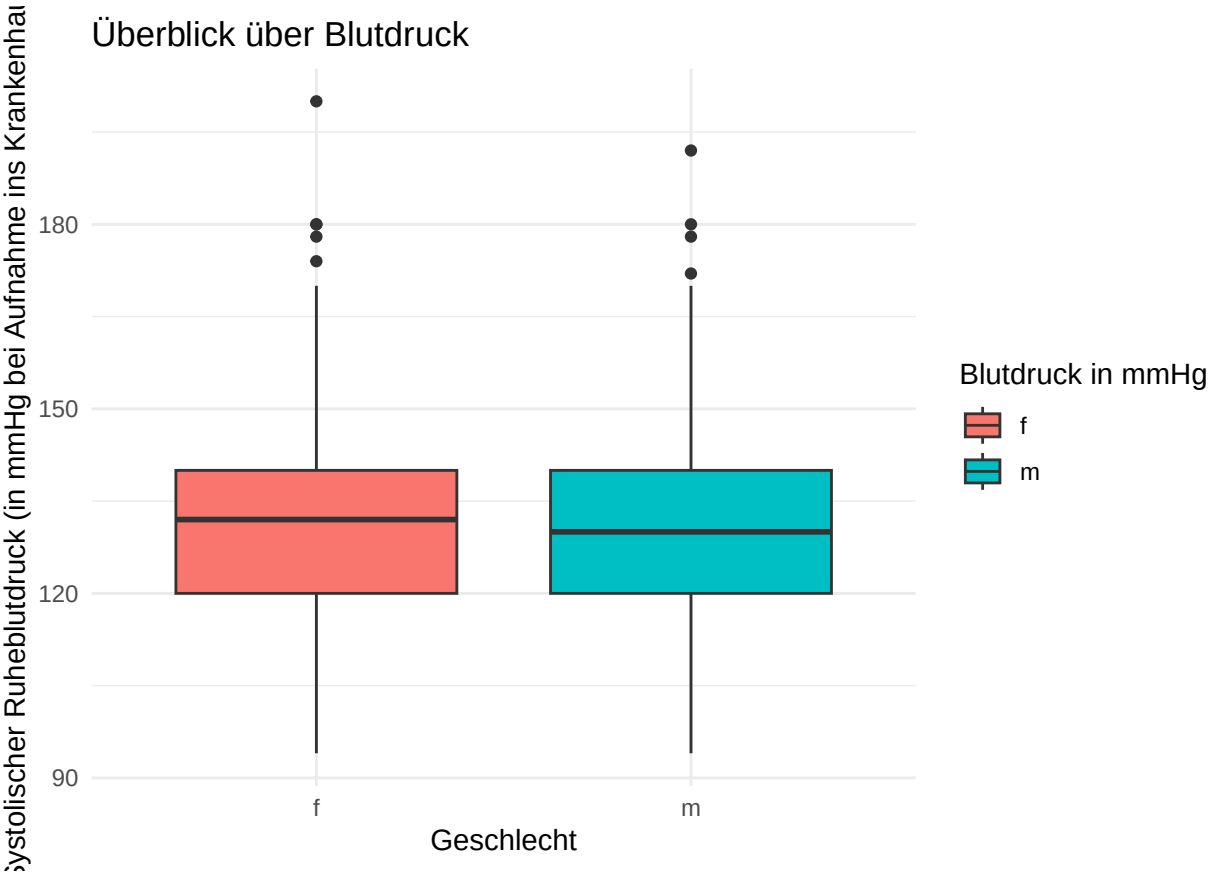
Systolischer Ruheblutdruck

Der systolische Blutdruck liegt beim gesunden Menschen bei ca. 120 mmHg.

Überblick über Blutdruck

Plot Verschaffen Sie sich einen Überblick, indem Sie ein Boxplot für den Blutdruck in Ruhe gruppiert nach dem Geschlecht plotten.

```
ggplot(herz, aes(x = sex, y = trestbps, fill = sex)) +
  geom_boxplot() +
  labs(title = "Überblick über Blutdruck",
       x = "Geschlecht",
       y = "Systolischer Ruheblutdruck (in mmHg bei Aufnahme ins Krankenhaus)") +
  scale_fill_discrete(name = "Blutdruck in mmHg") +
  theme_minimal()
```



Normalverteilt? Kann überhaupt davon ausgegangen werden, dass die Daten normalverteilt sind?

Konfidenzintervalle nach Erkrankung

Berechnen Sie die Konfidenzintervalle für den Ruheblutdruck aufgeschlüsselt nach der Diagnose (erkrankt/nicht erkrankt) und geben Sie das Ergebnis als `kable`-Tabelle aus:

```
konf_blutdruck_gesund = t.test(herz[herz$goal == "gesund", ]$trestbps, conf.level = 0.95)$conf.int
konf_blutdruck_krank = t.test(herz[herz$goal == "krank", ]$trestbps, conf.level = 0.95)$conf.int

result_table_blutdruck = data.frame(Diagnose = c("Gesund","Krank"),
```

```
                                        Untere_Grenze = c(konf_blutdruck_gesund[1], konf_blutdruck_kra
                                        Obere_Grenze = c(konf_blutdruck_gesund[2], konf_blutdruck_kran
                                        )
kable(result_table_blutdruck)
```

Diagnose	Untere_Grenze	Obere_Grenze
Gesund	126.7514	131.7486
Krank	131.4205	137.7161

Test, ob Kranke höheren Ruhe-Blutdruck haben

Überprüfen Sie mit einem Hypothesen-Test, ob Erkrankte einen höheren Ruhe-Blutdruck haben als gesunde Probanden.

Wie lauten die Hypothesen? Formulieren Sie die Hypothesen (H_0 und H_1).

Testanwendung Wenden Sie den Test mit R an. Was ist das Ergebnis?

```
# H0 Kranke haben einen niedrigeren oder gleichen Blutdruck wie Gesunde
# H1 Kranke haben einen höheren BLutdruck als Gesunde

test_blutdruck = t.test(herz$trestbps[herz$goal == "krank"],
                        herz$trestbps[herz$goal == "gesund"],
                        alternative = "greater"
                        ) # betrachten das Blutdruck bei Kranken höher ist als bei Gesunden

print(test_diagnose)
```

```
##
##  Welch Two Sample t-test
##
## data:  herz$chol[herz$goal == "krank"] and herz$chol[herz$goal == "gesund"]
## t = 1.4924, df = 298.64, p-value = 0.06832
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.9327489        Inf
## sample estimates:
## mean of x mean of y
##  251.4748  242.6402
```